Lessons Learned from Data Preparation for Geographic Information Systems using Open Data

Jun Iio
Dept. of Socio-informatics, Faculty of Letters, Chuo University
Hachioji-shi, Tokyo, JAPAN
iiojun@tamacc.chuo-u.ac.jp

ABSTRACT
The use of geographic information systems (GISs) has become widespread in data-driven industries, and they are utilized to visualize various kinds of spatial data using mappings. In addition to the large amount of available open-source GIS software, various types of data (e.g., boundary data for administrative regions, numerical data for individual areas, and data representing the objects on a map) are provided by local governments as open data. However, in many cases, the data have been inadequately maintained. Thus, advance preparation is required to utilize the data effectively. This paper discusses the work required to utilize open data by considering the case studies of Hachioji-city, Tokyo, Japan.

ACM Classification Keywords
H.2.8. Database Applications: Spatial databases and GIS

Author Keywords
Geographic information system; open data; boundary; cleansing; name identification.

INTRODUCTION
With the emergence of the digital era, geographic information systems (GISs) have become popular for visualizing and mapping spatial data. There have been several prominent software implementations, including Quantum GIS (QGIS) [3], GRASS GIS [8], MapServer, and other GIS tools [6, 10, 11]. These provide useful functions for visually analyzing spatial data, and GIS use is expected to spread because of the effectiveness of such functions.

Furthermore, GIS maps are realized as open source maps (OSMs) [7, 9], constituting an extremely large database with contributions from an enormous number of worldwide volunteers. The OSM makes it possible to freely utilize map data without any payments.

OpenSym '18, August 22–24, 2018, Paris, France

ISBN 978-1-4503-5936-8/18/08…$15.00

DOI: https://doi.org/10.1145/3233391.3233525

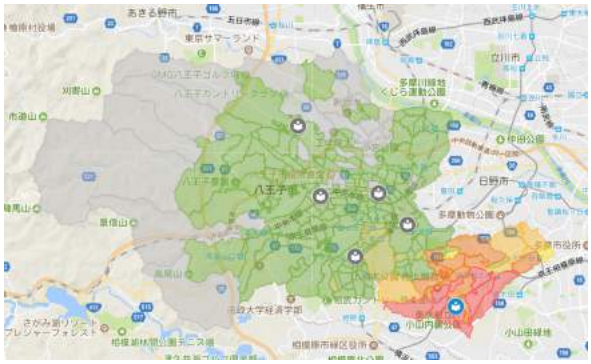

Figure 1. Library locations are shown on this map with six icons, and users are categorized according to the color of each associated area.

Additionally, the data to be visualized are provided by local governments as open data. However, the current state of data publication is far from the five-star rating that Tim Berners-Lee introduced in 2014 [5], and in many cases, these data are simply published on the internet without any consideration given to their format or effective reuse.

Background
Since July 2017, we have conducted collaborative research with the administrative sections of the Hachioji-city library. Hachioji-city is located at the west end of the Tokyo metropolitan area and is a mid-sized city with a population of approximately 600,000.

The Hachioji-city library collects various kinds of data during its daily administrative work, including information about users who check out books, what books are checked out, user addresses, and the average duration for which the books are checked out. It would be expected that such data could be used to make informed suggestions for improving library operations. However, library administrators do not have enough time to analyze such data. Therefore, a team was formulated of individuals from the Hachioji-city library and Chuo University to analyze the data collected from the daily activities of the library.

We began our research with a spatial analysis to visualize several types of data, such as the number of users and their registration rates, and from the Hachioji-city library by using GISs. Figure 1 illustrates the locations of the libraries

Figure 2. Hachioji-city local government publishes its data sets at "Hachioji-city Open-data Catalogue Page," which can be freely accessed via the internet.

in Hachioji-city and the distribution of library users. The six icons on the map indicates the locations of the libraries. The blue icon indicates the Minami-Osawa branch, and the distribution consists of users registered at that branch. The number of users in each area is shown by the color used for that area, where red indicates more than a thousand users, orange indicates over five hundred, green shows less than a hundred users, and gray means that there are no users. This map clearly illustrates a trend of registered library branch users.

Open-Data Available Online

To create the maps shown in figure 1, two types of data were required. One was a dataset used to define regional boundaries. Another was a dataset for each area, usually showing the distribution over all areas.

The former is represented by a set of points comprising single or multiple parts of polygons. In the case of Japan, if users want city and/or prefecture-level boundaries, they can obtain the data from the Ministry of Land, Infrastructure, Transport, and Tourism (MLIT). If smaller administrative unit boundaries are required, census-related data provided by the Ministry of Internal Affairs and Communications (MIC) are available. The latter are accessible online only if the local government has published it as open data.

Recently, the open data movement is rapidly gaining popularity globally, allowing people to access enormous quantities of data from local governments. As an example, Figure 2 shows that Hachioji-city publishes various types of open data collected via their daily administrative work.

PROBLEMS

As described in the previous section, both types of data, representing the boundaries and target to be analyzed, can be freely obtained as administrative services from the internet. However, some problems remain with both types of data. Thus, some modifications are necessary to successfully conduct our geospatial research.

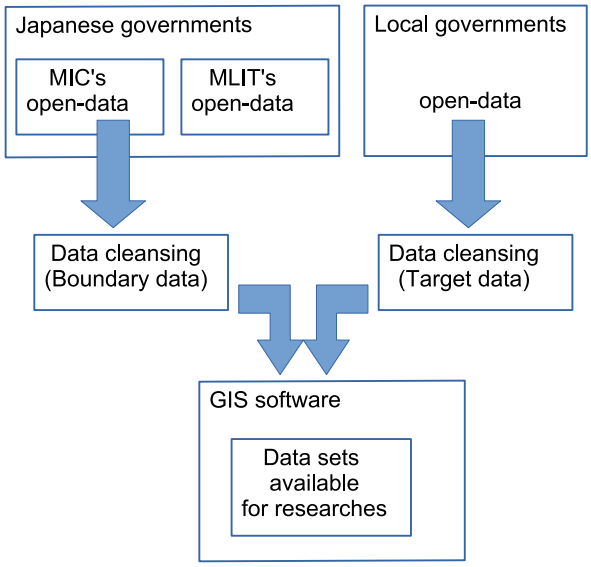

Figure 3. Scheme to utilize appropriate data sets to conduct our research.

Figure 4. Map of Kanagawa prefecture with boundaries indicating shapes of cities and towns.

Figure 3 illustrates a scheme to utilize appropriate data sets to conduct our study. In this paper, the problems with boundary-data and open-data are described; then, an example of the solution to eliminate the defects in the boundary-data is presented.

Problems with Boundary-Data

There are two types of inappropriate situations in the boundary data from MLIT or MIC. If one uses the MLIT data, it should be noted that the dataset is very complicated.

Figure 4 is a map of Kanagawa prefecture with some boundaries showing the shapes of cities and towns. Cities and towns are clearly separated from each other by red lines, appearing adequate for spatial analysis.

However, the data contain very detailed shapes representing rock reefs and breakwaters. Figure 5 illustrates a typical ex-

Figure 5. Expanded map showing Manazuru town, located at the southwest end of Kanagawa prefecture and facing the sea. It contains detailed shapes that represent rock reefs and seawalls.

dataA.csv

```
name,age
八王子市横山町,43.64
八王子市八日町,39.8
八王子市八幡町,44.9
八王子市八木町,44.31
八王子市追分町,48.34
八王子市千人町1丁目,45.36
八王子市千人町2丁目,41.91
八王子市千人町3丁目,44.97
八王子市千人町4丁目,44.39
...
```

dataB.csv

```
name,age
八王子市横山町,43.64
八王子市八日町,39.8
八王子市八幡町,44.9
八王子市八木町,44.31
八王子市追分町,48.34
八王子市千人町一丁目,45.36
八王子市千人町二丁目,41.91
八王子市千人町三丁目,44.97
八王子市千人町四丁目,44.39
...
```

Figure 6. Differences between dataA.csv and dataB.csv. DataB.csv uses Chinese (i.e., Kanji) characters to represent the numbers of some districts, whereas dataA.csv uses Arabic numerals.

ample. Manazuru town is located at the southwest end of Kanagawa prefecture and faces the sea. It contains detailed shapes that represent rock reefs and seawalls.

Even if MLIT considers such details to be very important for representing the real topography of Japan, they are not required to analyze spatial data distributed over the entire region. Therefore, we should discard small portions (that may not be needed) and aggregate several areas into one larger region.

Problems with Open-Data

The open-data provided by local governments also have certain defects. Figure 6 shows trivial differences between two datasets represented in the comma-separated-value (CSV) format. Although the two datasets are basically, there are some differences in the notations of the names of a few regions.

This figure shows that Sen'nin-cho, Hachioji-city has four districts. In dataA.csv, these four districts are written using Arabic numerals as 1-Choume and 2-Choume. In contrast, the district numbers are written as Chinese numerals in dataB.csv. A person[1] can understand that dataA.csv and dataB.csv are the same. However, a computer is incapable of

[1]This would only be true for people who can read Japanese or Chinese characters.

```
cat dataA.csv | \
  awk 'BEGIN { FS="," } { \
    cmd = "echo "$1" | tr 1234567890 一二三四五六七八九〇"; \
    cmd | getline out; \
    printf "%s,%s\n",out,$2; \
    close(cmd) }' > dataB.csv
```

Figure 7. Example of a shell script to convert Arabic numerals to Chinese numerals in the name field.

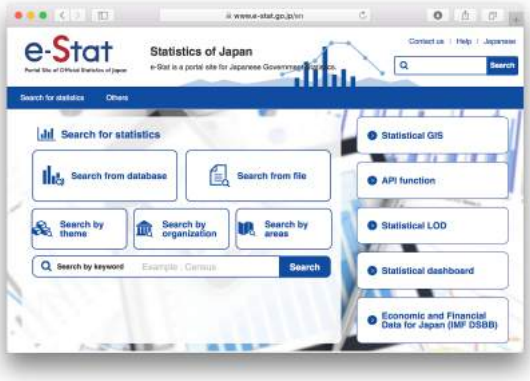

Figure 8. Homepage of e-Stat website. Anyone who wants to map the data for small administrative regions can download boundary data for the applicable scope of cities and towns.

determining that these two datasets are identical. Therefore, some preparations have to be conducted for name aggregation.

Figure 7 shows a shell script that converts Arabic numerals to Chinese numerals. In this script, the "tr" command is applied only within the first field. Thus, it converts the appropriate characters in the names of regions without conversion of the remaining fields.

EXAMPLE OF SOLVING A PROBLEM

This section discusses a case example of the preparation of boundary data. As previously described, anyone who wants to map data for small administrative regions can obtain boundary data from the "e-Stat" website operated by MIC in Japan (Figure 8.)

Unfortunately, some datasets provided by the MIC website may have numerous defects, because the boundary data are created to determine scope for researchers from a series of census investigations instead of precise administrative boundaries from local governments. Therefore, anyone who wants to use administrative boundary data for research must adjust the boundaries defined by e-Stat to obtain the precise administrative boundaries.

The case of Tate-machi, Hachioji-city, is shown in Figure 9. Tate-machi is represented by six subregions. However, it must be represented as only one region, because there are no administrative subregions.

There are three small areas inside the other regions, and it is very easy to remove these inner regions. These boundary data

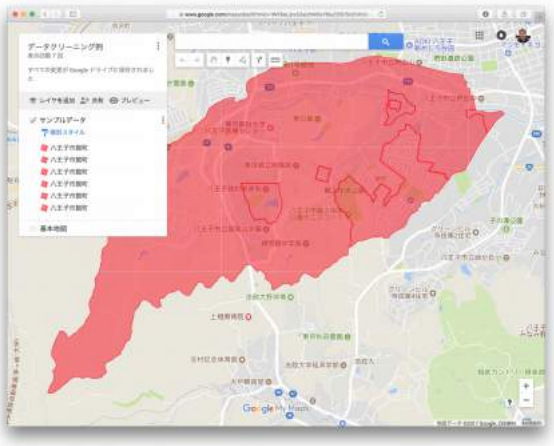

Figure 9. Example of boundary data showing Tate-machi in Hachioji-city. It contains six regions, although it must be represented as only one region.

Figure 10. To combine two regions, sequence of control points not included in the border are merged into one sequence of points.

are stored in a file with the GeoJSON[2] or KML[3] format [12]. It is necessary to remove these three parts and several subparts defined as inner boundaries by editing the data file.

Consolidating two side-by-side regions is more complicated than the previously mentioned case. Figure 10 provides a straightforward overview for merging two regions.

We assume that the boundary data contains a set of control points defined along the edges of the region in a counterclockwise order. Under this assumption, the sequence of control points consisting of the boundary of one region is divided into two parts. The first is the sequence of points on the border of the other region. The other is the sequence of points not included in the border. For each region, the latter control points are picked up, and combining them adequately results in a new region of the two combined regions.

After repeating the process of selecting two adjoining regions, combining them recursively, we obtain a unique shape that correctly represents one administrative region. Figure 11

[2]JSON stands for JavaScript Object Notation, and GeoJSON is the particular format to store geographical data in a JSON file.

[3]KML is the acronym for keyhole markup language. See **`https://developers.google.com/kml/`** for more information on the KML format for geographic data.

Figure 11. Three regions are merged into one region in two combining processes in sequence.

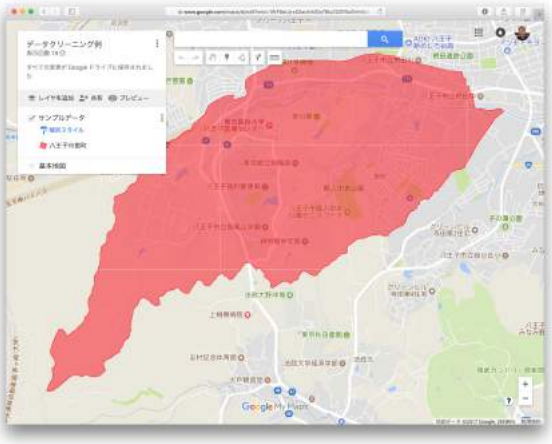

Figure 12. Appropriate boundary data for Tate-machi after the cleansing process has been completed.

illustrates the merging of three regions into one region using two sequential combining processes. The final result of a cleansing process is shown in Figure 12. Note that occasionally cases exist where the points used for the border between two regions are not precisely identical. Thus, we need to carefully check the distances between two corresponding points, and if they are critically close, they must be identified as the same points.

RELATED WORK

There have been many studies on the open-data strategies. Huijboom and Broek [4] examined the open-data strategies in several countries and discussed key features, barriers, and determinants for progress and effects. Conradie and Choenni [1] argued the barriers for local government releasing their data as open-data. Veljković *et al.* [13] proposed a benchmark for the open government and its application from the open data perspective. However, these studies are based on the governments' perspective, instead of the citizens' perspective. That is, their discussions do not focus on how the data are practically reused. This study intends to fill the gap between the theoretical consideration on the open-data and its practical use.

There have also been many investigations on the reuse of GIS data. Halfawy *et al.* [2] discussed the main requirements for the standard data models of GIS data and the importance of interoperability from an asset management perspective. Zhang *et al.* [14] described the importance of the

geography markup language (GML), as the standard data exchange format from the view of web-oriented system design, where many information systems are currently provided as the web systems.

Obviously, these standard models and data exchange formats are important. However, it is not sufficient to only prepare these standards. The content of the data itself should also be appropriately managed, including the granularity of boundary data, as previously mentioned herein.

CONCLUSIONS AND FUTURE WORK

In Japan, some ministries (MLIT and MIC) have published boundary data for defining administrative regions on the internet, and such data can be freely downloaded at no cost. However, the raw data provided are inappropriate for use in spatial analyses using GIS tools. This paper reported procedures required to clean the boundary data provided by the Japanese government as open-data. Additionally, various data have been made available as open-data by local governments. They also have room for modification to be effectively used in research activities.

The methods explained in this paper, including the name aggregation and data cleansing, are operated by applying some scripts manually. Therefore, a trial to implement a system to automatically realize these procedures remains in the scope of future work.

REFERENCES

1. Peter Conradie and Sunil Choenni. 2014. On the barriers for local government releasing open data. *Government Information Quarterly* 31 (2014), S10–S17. `http://www.sciencedirect.com/science/article/pii/S0740624X14000513` ICEGOV 2012 Supplement.
2. Mahmoud R. Halfawy, Dana J. Vanier, and Thomas M. Froese. 2006. Standard data models for interoperability of municipal infrastructure asset management systems. *Canadian Journal of Civil Engineering* 33, 12 (2006), 1459–1469.
3. Marco Hugentobler. 2008. *Quantum GIS*. Springer US, Boston, MA, 935–939.
4. Noor Huijboom and Tijs Van den Broek. 2011. Open data: an international comparison of strategies. *European journal of ePractice* 12, 1 (2011), 4–16.
5. Krzysztof Janowicz, Pascal Hitzler, Benjamin Adams, Dave Kolas, II Vardeman, and others. 2014. Five stars of linked data vocabulary use. *Semantic Web* 5, 3 (2014), 173–176.
6. Tyler Mitchell. 2005. An Introduction to Open Source Geospatial Tools. (2005). `http://www.oreillynet.com/pub/a/network/2005/06/10/osgeospatial.html`
7. Pascal Neis and Dennis Zielstra. 2014. Recent Developments and Future Trends in Volunteered Geographic Information Research: The Case of OpenStreetMap. *Future Internet* 6, 1 (2014), 76–106.
8. Markus Neteler, M. Hamish Bowman, Martin Landa, and Markus Metz. 2012. GRASS GIS: A multi-purpose open source GIS. *Environmental Modelling & Software* 31 (2012), 124–130.
9. Leysia Palen, Robert Soden, T. Jennings Anderson, and Mario Barrenechea. 2015. Success & Scale in a Data-Producing Organization: The Socio-Technical Evolution of OpenStreetMap in Response to Humanitarian Events. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 4113–4122.
10. Stefan Steiniger and Andrew J. S. Hunter. 2012. *Free and Open Source GIS Software for Building a Spatial Data Infrastructure*. Springer, Berlin, Heidelberg, 247–259.
11. Stefan Steiniger and Andrew J. S. Hunter. 2013. The 2012 free and open source GIS software map – A guide to facilitate research, development, and adoption. *Computers, Environment and Urban Systems* 39 (2013), 136–150.
12. Stefan Steiniger and Andrew J. S. Hunter. 2016. *Data Structure, spatial data on the web*. Wiley.
13. Nataša Veljković, Sanja Bogdanović-Dinić, and Leonid Stoimenov. 2014. Benchmarking open government: An open data perspective. *Government Information Quarterly* 31 (2014), 278–290.
14. Chuanrong Zhang, Tian Zhao, and Weidong Li. 2015. *Geospatial Data Interoperability, Geography Markup Language (GML), Scalable Vector Graphics (SVG), and Geospatial Web Services*. Springer, Cham.